Considerations of IP Multicast for Load Balancing in Proxy Mobile IPv6 Networks

Tien-Thinh Nguyen[a,1,*], Christian Bonnet[a]

[a] *Department of Mobile Communications*
EURECOM
450 Route des Chappes, 06410 Biot, France

Abstract

Proxy Mobile IPv6 (PMIPv6), taking advantages of the network-based mobility management, enables mobility support for the mobile nodes (MNs) without requiring their involvement in mobility signaling. However, as a centralized mobility management, PMIPv6 relies on a central mobility entity, i.e., Local Mobility Anchor (LMA) to provide the mobility support. The LMA is responsible for maintaining the mobile node's (MN) reachability state and forwarding the traffic from/to the current location of the MN. As mobile traffic demand rapidly increases, it is easy to make the LMA a bottleneck and a single point of failure. Therefore, load balancing (LB) mechanism among LMAs is a promising solution for these issues. Although previous studies proposed several solutions for distributing the load among the LMAs, none of them considers the multicast service. From the fact that the multicast service is expected to be widely used for delivering multimedia traffic (which will account for the majority of mobile traffic), it can also be considered as a crucial load factor. As a result, the efficiency of the existing solutions may be degraded when considering multicast. Furthermore, applying the existing LB mechanisms can raise several issues for not only the ongoing unicast sessions but also the multicast ones. To tackle these issues, this paper proposes a new LB solution which mainly focuses on the multicast service. The experiments and the numerical results show that this solution helps to better distribute the load among the LMAs while greatly reducing the multicast service disruption as well as avoiding the influence on the ongoing unicast sessions. In addition, the proposed solution can co-operate with the existing proposals to improve the performance of the network.

Keywords: Load Balancing, Multicast-based Load Balancing, Proxy Mobile IPv6, IP Multicast, Multicast Listener Mobility.

1. Introduction

Nowadays, the mobile data services have become an essential part of many consumers' life [1, 2]. So far, users are using their mobile devices (e.g., smartphones and tablets) not only for personal life (e.g., making voice/video calls, sending email, watching video/TV, playing online games, and so on) but also for work (general and job-specific work applications such as multimedia conferencing, and distance learning, etc.) on a regular basis [3, 4, 5]. As a result, the mobile data traffic has been nearly doubled each year during the last few years [6]. This trend is expected to continue in upcoming years, especially with the deployment of 4G networks. The increase in traffic is mainly driven by mobile video traffic: estimates say that the mobile video traffic will account for 66.5 percent of total data traffic by 2017 [1]. The wide usage of mobile data services has been driven by the variety of different reasons such as: the increasing number of mobile devices which become more and more powerful and intelligent, the enhancement of wireless access technology in terms of coverage, speed and quality, as well as the explosion of mobile applications [6]. The mobility of the devices puts a new requirement on mobile operators to provide connectivity anywhere and at

*Corresponding author
Email addresses: tien-thinh.nguyen@eurecom.fr (Tien-Thinh Nguyen), christian.bonnet@eurecom.fr (Christian Bonnet)
[1] Phone number: (+33) 4 93 00 82 15

anytime. Moreover, providing consistent and continuous seamless services is required for satisfying user's expectations and fulfilling even the high application requirements in terms of service disruption on the move [7].

In this context, various IP mobility management protocols have been introduced by the Internet Engineering Task Force (IETF)[2] ranging from the host-based (including Mobile IPv6 (MIPv6) [8] and its extensions e.g., Hierarchical Mobile IPv6 (HMIPv6) [9]) to the network-based mobility approach (e.g., Proxy Mobile IPv6 (PMIPv6) [10] and Fast Proxy Mobile IPv6 (FPMIPv6) [11]). On one hand, PMIPv6, as a network-based mobility management, provides mobility for the mobile nodes (MNs) without their involvement. This means the network handles the mobility management on behalf of the MN. As a result, PMIP helps to avoid the complexity of the protocol stack in the MN as well as to reduce tunneling overhead (over the air) and handover latency compared to the MIPv6. It is achieved by introducing two new network entities, namely the Local Mobility Anchor (LMA) and the Mobile Access Gateway (MAG). The former, similar to the home agent (HA) in MIPv6, is in charge of tracking the location of the MN and redirecting the MN's traffic towards its current topological location. While the latter is responsible for detecting and registering the movement of the MN. On the other hand, both PMIPv6 and MIPv6 are the centralized mobility approaches, which rely on the mobility anchor to enable mobility support (LMA in PMIPv6 and HA in MIPv6). In PMIPv6, it is common to have a huge number of devices associated with the LMA. As the traffic demand is increased rapidly [1], a traffic bottleneck can be formed at the LMA. Consequently, the quality of the ongoing sessions could be degraded (e.g., longer queuing delay, and increased packet loss). Also, in a heavy load condition, the LMA can drop the new sessions. In this circumstance, mobile network operators may need to deploy multiple LMAs in a large PMIPv6 domain, so that the traffic can be distributed among the LMAs [10]. Yet, it is highly possible that some LMAs become overloaded while the others are underutilized. Thus, load balancing (LB) among the LMAs is needed.

Several LB proposals [12, 13, 14, 15] have been introduced to allow the LMA to be dynamically assigned and changed according to the load of all LMAs in the domain. When an MN initially attaches to the domain, the LB will be executed to select the appropriate LMA in terms of load to serve this MN (namely proactive-MN approach). However, the varying session rate (of the existing MNs) and data rate (of the existing sessions) may cause load-unbalanced situation between the LMAs. In order to address this issue, the LB can be triggered when the load of an LMA exceeds a specified threshold (called reactive-MN approach). In this case, an MN will be selected to move from the overloaded LMA to a less loaded one. Yet, changing LMA causes some issues for the ongoing sessions such as service disruption and packet loss.

As Internet is widely deployed and spread across a large area, it carries a variety of common information resources and services. In a sharing world, the group communication service, which refers to the ability to send data to several receivers at the same time, is naturally becoming more and more important especially in some areas like multimedia distribution, gaming, and financial services, etc [16]. In this context, the scalability and bandwidth efficiency from the multicast routing make the IP multicast a remarkable solution from the application point of view to allow the mobile networks to deal with a huge number of traffic, particularly, in mobile environments where users usually share frequency bands and limited capacity [17]. However, its role has been neglected in all existing LB proposals. As such, the consideration of multicast in the existing LB mechanisms can bring several issues from both load balancing (efficiency degradation) and multicast service perspective (e.g., tunnel convergence problem [18] and service disruption).

For these reasons, a LB mechanism which takes the multicast service into account is needed. In this paper, we will introduce such a LB mechanism, the so-called multicast-based mechanism. The key idea is that by separating the multicast LB from the unicast LB, the proposed solution helps better distribute the load among the LMAs in runtime, thus, improving the efficiency of resource utilization. In more details, when an LMA is overloaded, a multicast session will be selected to move to a less loaded one. The LB will also be executed when a listener starts a new multicast session to select the appropriate LMA to serve this session. As a result, the proposed solution does not influence the ongoing unicast/multicast sessions (except the selected session with which the multicast service disruption, in

[2]IETF, http://www.ietf.org

most cases, satisfies the requirements for the real-time services [19]).

As this article is an extension of [20], we will make a quick view on the issues caused by applying multicast in the existing proposals as well as the multicast-based solution. We will discuss in detail the criteria for the selection of the LMA and the multicast session. Next, the performance analysis will be done regarding the LMA load and the multicast service disruption. Finally, we will evaluate the multicast-based solution in terms of load distribution among LMAs using a near-to-real testbed. The testbed which is a combination of virtual machines and the network simulator NS-3 [21] has been deployed to reduce the hardware cost and to provide more flexible experiment while allowing to obtain the realistic results. It is noted that this paper mainly focuses on the multicast listener.

The rest of this paper is organized as follows. Section 2 presents the existing LB mechansims as well as the issues when considering multicast with these mechanisms. Section 3 introduces the multicast-based LB as well as the criteria for the LMA and multicast session selection. Section 4 presents the performance analysis regarding LMA load and multicast service disruption. Section 5 takes a look on the experiment testbed including the testbed description, the experiment scenarios and the collected results. Section 6 discusses the limitations of the proposed solution as well as security consideration. Finally, Section 7 concludes the paper and provides perspectives for future works.

2. Related Work

2.1. Load Balancing: from Mobile IPv6 to Proxy Mobile IPv6

In fact, both MIPv6 and PMIPv6 are two typical examples of the centralized mobility management protocol. They rely on a central mobility anchor e.g., HA in MIPv6 and LMA in PMIPv6 to provide mobility support. The central mobility anchor is responsible for maintaining the mobility signaling and data traffic of all connected MNs in the domain. As a result, it raises the bottleneck and single point of failure. Therefore, the load balancing mechanism among the mobility anchors is required.

In this context, several proposals have been introduced to balance the load among different HAs in MIPv6 and different LMAs in PMIPv6. For example, in MIPv6, Home Agent Load Sharing architecture [22] was presented to select the most appropriate HA during the bootstrap phase. This solution relies on a set of parameters including the number of active associated MNs, current bandwith consumption of HA as well as the location of the HA. In [23], the authors proposed a method to allow reassigning the HA whenever a timer expires. However, this solution does not take into account the service disruption problem caused by the changing of the HA. In [24], the authors introduced the "Virtual HA Reliability Protocol" which is an extension to MIPv6 to support load balancing among HAs. However, as a LB solution for MIPv6, all above mentioned proposals typically require the participation of the MN into the load balancing-related signaling. Therefore, they cannot be directly applied in a PMIP environment.

Regarding the LB mechanism in a PMIP domain, there are two main strategies: LB among the LMAs [12, 14, 13, 15] and LB among the MAGs [25, 26, 27]. It is noted that a typical PMIPv6 deployment allows one LMA to serve approximately 250 MAGs [28]. Moreover, since the LMA plays the role of a mobility anchor, changing LMA can cause some issues to the ongoing sessions e.g., service disruption and packet loss. For these reasons, in this paper we focus on the LB mechanism among the LMAs.

There are two main approaches for LB among LMAs in PMIPv6, namely, proactive-MN and reactive-MN. In the proactive-MN approach [12, 14], the LB will be executed in the initial phase of an MN to select the least loaded LMA. In other words, when an MN initially attaches to a PMIPv6 domain, the least loaded LMA will be selected to serve this MN in runtime based on the current load of all LMAs in the domain. All mobility sessions of this MN then would be anchored at the assigned LMA during their lifetime in the domain. The main advantage of this approach is that it does not influence the ongoing sessions of the registered MNs. However, since it is executed in the initial phase of an MN, the varying session rate and data rate may cause the unfair load distribution among the LMAs. In the reactive-MN approach [13, 15], the LB will be triggered when the LMA load exceeds a specified threshold. The overloaded LMA will select one (or several) MN(s) to move to a less loaded LMA (called target LMA, or tLMA). This approach allows the network to adapt to the current situation. Thus, it may give a better performance e.g., distributing load among LMAs and increasing the

reliability. Since the LMA plays the role of the mobility anchor for the MN, changing LMA during the mobility session could impact the selected MNs ongoing sessions in terms of service disruption and packet loss. For this reason, this change is not recommended by the IETF [12, 14]. In addition, the existing proposals only consider the ongoing sessions as the unicast ones. In more details, in [13], the load information of all LMAs can be collected and managed at the authentication, authorization and accounting (AAA) server which then selects the tLMA among the LMAs in the domain. In [15], the authors do not mention about how to select and get the address of the tLMA. The MN selection can be based on some policies such as: i) The MN having a real-time service should not be selected [13]; and/or ii) The MN with the highest session-to-mobility ratio should be selected [15].

In conclusion, the existing load balancing mechanisms can be considered as a per-MN approach. That means these proposals are based on the assignment of the MN with an anchor. All flows of this MN are anchored at the assignment anchor during their lifetime. Moreover, these mechanisms rely on a central entity to collect and manage the collected load of the mobility anchors. As can be seen later in the experiments, the load mainly depends on the traffic, rather than the number of associated MNs. Therefore, a per-flow load balancing mechanism should be provided. In addition, the existing proposals only consider the ongoing sessions as the unicast ones. How the LB works with the multicast is still an open question. It is also necessary to consider IP multicast to avoid the potential impact of multicast service on the efficiency of load balancing.

2.2. Multicast Mobility Support in PMIPv6

As described in the base solution [28], to support multicast in a PMIPv6 domain, the multicast router (MR) function (executing a multicast routing protocol e.g., Protocol Independent Multicast - Sparse Mode (PIM-SM) [29]) and the Multicast Listener Discovery (MLD) proxy function [30] need to be deployed at the LMA and the MAG, respectively [28]. In this case, a listener always receives the multicast traffic from its LMA via the LMA-MAG tunnel, just like the unicast traffic.

2.3. Multicast Considerations with the Existing Load Balancing Mechanisms

Regarding the proactive-MN approach [12, 14], this approach only takes the current load of the LMAs (neither unicast nor multicast service) into account. As a result, the varying session and data rate of the registered MNs may result in an unfair load distribution between LMAs. On the other hand, the reactive-MN approach [13, 15] allows selecting one MN to move from an overloaded LMA to a less loaded one (called target LMA, or tLMA) (see Fig. 1). The Proxy Binding Update (PBU)/Proxy Binding Acknowledgment (PBA) messages are then exchanged between the current LMA (cLMA) and the tLMA allowing the tLMA to serve as the new mobility anchor of the MN. When considering multicast in the reactive-MN approach, if there is more than one listener (including the selected one) associated with the cLMA and subscribing to the multicast channel, the cLMA will continue forwarding this channel. Consequently, moving the MN cannot help significantly reduce the LMA load, especially when the load generated by this MN is mainly derived from this channel. The total load of all LMAs may also be increased since the tLMA may need to join the channel. Also, from the multicast service point of view, several procedures (e.g., obtaining the MN's subscription information via the MLD Query/Report process, joining the multicast delivery tree) need to be executed in order to allow the MAG to continue receiving the traffic (from the tLMA). As a result, it experiences a noticeable service disruption for the ongoing multicast channels. Additional mechanisms (e.g., MLD proxy peering function [31]) are required to reduce the service disruption time. In the worst case scenario, as the LMA selection algorithm does not take multicast service into account, the tLMA may not support the multicast capability. In other words, the multicast service cannot be guaranteed at the tLMA. Also, since many proxy instances are installed at MAG, it may cause the tunnel convergence problem [18].

3. Multicast-based Load Balancing Solution

In this section, at first, some criteria to select the appropriate LMA and multicast session for the LB purpose will be discussed. Two different approaches of the multicast-based solution i.e., the proactive-multicast (or MAG-initiated) and the reactive-multicast (or LMA-initiated) are then considered. In the former case, LB will be invoked when an MN starts a new multicast session to select a suitable LMA to serve this session. In the latter case, LB will be executed whenever an LMA is

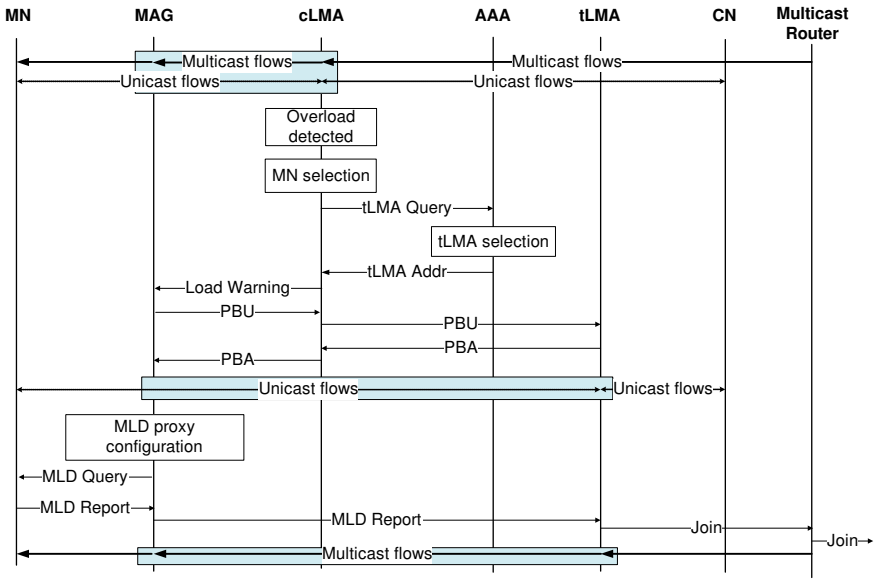


Figure 1: IP Multicast in the reactive-MN approach.

overloaded by selecting a multicast session to move from this LMA to the less loaded one. It can be done thanks to an extension to MLD proxy supporting multiple upstream interfaces [32]. In this case, only one proxy instance is deployed at MAG with multiple upstream interfaces towards different LMAs. Consequently, the MN can receive the multicast traffic from a less loaded LMA, while obtaining the unicast traffic from its LMA.

3.1. LMA and Multicast Session Selection

3.1.1. Target LMA Selection

Target LMA selection is first based on the channel policy which is defined by the operators (if exist). Otherwise, the LMA selection relies on the following policies (from high to low priority): i) The least loaded LMA among the (not overloaded) LMAs having the multicast forwarding state for this channel should be selected; and ii) The LMA with the lowest load in the domain should be selected. The selection policies come from the fact that if the channel is already available at the selected LMA (tLMA) with a negligible increase of load, the tLMA can forward this channel to the MAG [16].

To do so, a new logical entity, the so-called load balancing controller (LBC), has been introduced. This entity collects and manages the load state information of all LMAs as well as their active multicast channels in the domain. It is also responsible for the LMA selection. Upon the location of the LBC, three different schemes can be considered as below:

- Centralized LBC entity (CE-LBC): The functionality of the LBC is responsible by a central entity, called C-LBC. This entity is similar to the notion of rfLMA as described in [12]. The LMAs periodically report their current load and their active multicast channels to the C-LBC by using an extension to the PBU/PBA message with the load information [12]. The C-LBC can be co-located with the AAA server.

- Distributed LBC function on the LMAs (DLMA-LBC): The LBC function is executed in a distributed manner among the LMAs. Each LMA maintains a so-called Load Table which includes load information of all LMAs and their associated multicast channels in the PMIPv6 domain. Each LMA periodically exchanges the information with each other in the domain, for example, by setting a common multicast group for all LMAs. Based on the load information in the Load Table, the LMA can select the most appropriate LMA candidate.

- Distributed LBC function on the MAGs (DMAG-LBC): In this case, the load and list of the current multicast channels of all LMAs is

collected and stored at the MAGs. The MAG can obtain these information of the associated LMA by using an extension of PBU/PBA messages or an extension of the Heartbeat message with the load information [33].

Without loss of generality, this paper only considers the first scheme. As all LMAs periodically report their workload to the C-LBC, the frequency of the workload report should be carefully examined as the trade-off between the precision of the load state and the signaling/processing overhead. One possible solution is that the LMA only reports its workload when its load exceeds/is lower than a certain load level. For example, the load of an LMA can be devided into three levels: low (e.g., the load is less than 30% of capacity), medium (from 30% to 70%), and high (above 70%). Therefore, the LMA only sends its load to the C-LBC whenever its load level changes. The higher number of levels can provide more accurate load state of an LMA at a cost of an increase in the generated traffic. Besides, the load threshold value should also be considered. The threshold value of each LMA depends on its capacity e.g., it may be set to a certain percent of the capacity. The threshold value can also be dynamically calculated upon the current load of all LMAs in the domain such as the average of the current load of all LMAs. Note that if the threshold value is dynamically calculated, additional signalling overhead can be introduced. In this paper, the threshold value selection is left for the implementation works.

3.1.2. Multicast Session Selection

The multicast session can be selected following some criteria: i) To reduce the potential impact on the ongoing sessions, the real-time and delay-sensitive sessions should not be selected. However, if all sessions are the real-time and delay-sensitive ones, the session with the highest data rate should be selected; and ii) The session requiring the highest data rate with the smallest number of subscribed listeners should be selected. It is noted that to better select LMA, the LMA selection algorithm should take the expected load of the selected multicast session into account.

3.2. Load Balancing in the Proactive-Multicast Approach

The signaling procedure for the proactive-multicast (MAG-initiated) approach is illustrated in Fig. 2. When a registered MN wishes to subscribe to a multicast channel and this channel is available at the current MAG, the MAG will forward it directly to the MN. Otherwise, it will contact the C-LBC to get the address of an LMA (following the criteria as stated earlier), which can be served as the multicast anchor point for this session. After joining the channel via the tLMA, the MAG can receive the multicast packets and forwards them to the MN. Note that the communication between the MAG and the C-LBC can be done by extending the Remote Authentication Dial In User Service (RADIUS) protocol for PMIPv6 [34] or PBU/PBA messages. Regarding the distributed scheme, for example, the DLMA-LBC (where the functionality of LBC is distributed among LMAs) the procedures are almost similar to those in the CE-LBC scheme. However, the MAG, instead of contacting with the C-LBC, will ask the MN's LMA to get the address of the target LMA. That means the MN's LMA plays a similar role as the C-LBC.

3.3. Load Balancing in the Reactive-Multicast Approach

Fig. 3 shows the signaling procedure for the reactive-multicast (LMA-initiated) approach. When an LMA (cLMA) is overloaded (its load exceeds a certain threshold), a multicast session will be selected to move from this LMA to a less loaded one (tLMA). After obtaining the tLMA address from the C-LBC, the cLMA sends the tLMA's address and the selected multicast session information to all related MAGs via a load-warning message (e.g., using an extension to the Update Notification message (UNP) [35]). The C-LBC also requests the tLMA to join the channel in advance to reduce the multicast service disruption. The MAG then sends an MLD Report to the tLMA to join the channel. Afterwards, the MAG can receive the multicast packets from the tLMA instead from the cLMA. In the meantime, the cLMA can leave this channel in order to lower its load. It is noted that in case of DLMA-LBC, the cLMA can decide itself on the tLMA based on its Load Table.

3.4. Handover Consideration

As can be seen in Fig. 4, if the MN performs a handover between two MAGs, the normal PMIP operation will be executed to update the routing information at the MN's LMA and the new MAG. Then, the similar process as for the new multicast

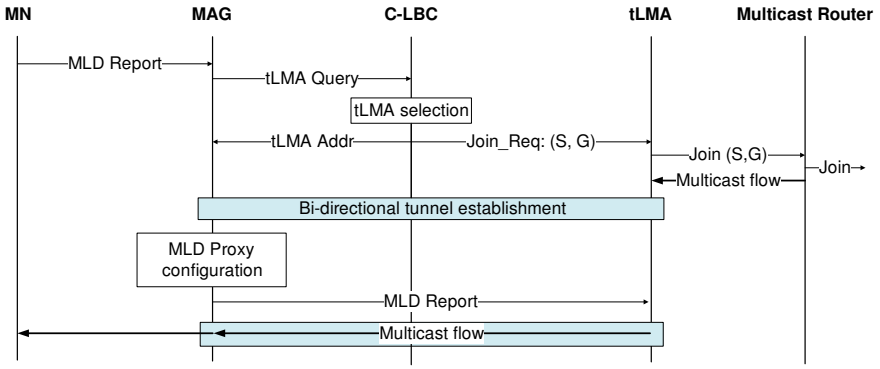


Figure 2: Proactive-multicast approach.

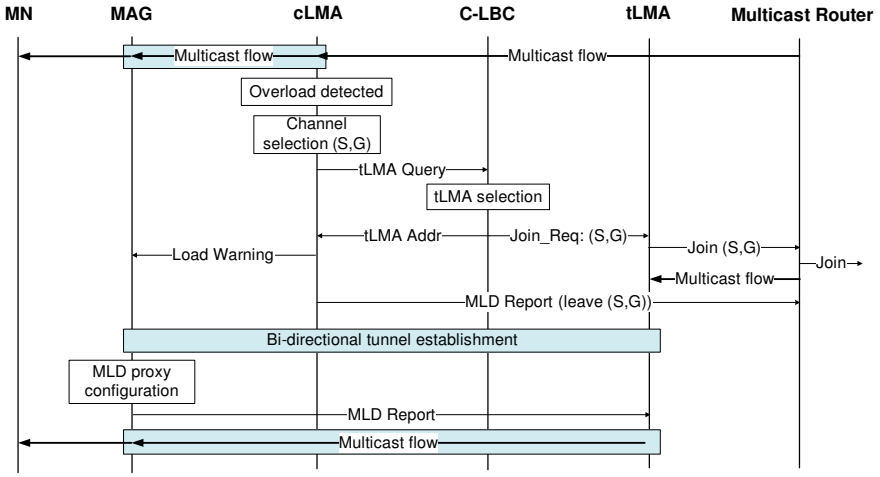


Figure 3: Reactive-multicast approach.

session at the new MAG will be undertaken to select the appropriate LMAs to serve the ongoing multicast channels.

4. Performance Analysis

In this section, at first, we will highlight the different load factors imposed on the LMA. Based on that the comparison will be conducted between the reactive-MN and the reactive-multicast approach regarding their efficiency. The multicast service disruption time will also be considered.

4.1. Load Analysis

4.1.1. Load imposed on LMA

As stated previously, to support multicast in a PMIPv6 domain, the multicast router (MR) function and the MLD proxy function [30] need to be deployed at the LMA and the MAG, respectively. All multicast traffic passes through the MAG-LMA tunnel, accordingly. As such, the load of the LMA comes from two main parts: the load from the typical LMA's tasks (L_{lma}^{lma}) and the load from the MR's tasks (L_{lma}^{mr}). It is noted that a minor amount of load which is imposed by the background processes (e.g., system processes) is ignored in our analysis. Thus, we have

$$L_{lma}^{(.)} = L_{lma}^{lma} + L_{lma}^{mr}. \quad (1)$$

As a typical LMA, it performs three main logic functions: mobility routing (processing the unicast traffic from/to the associated MNs), location management (processing PBU/PBA, updating binding cache, maintaining tunnel, etc.) and home network prefix (HNP) allocation [10]. As a result, L_{lma}^{lma} comes from three main parts L_{lma}^{mor}, L_{lma}^{lm}, and L_{lma}^{hal} corresponding to these logic functions. L_{lma}^{mor} and

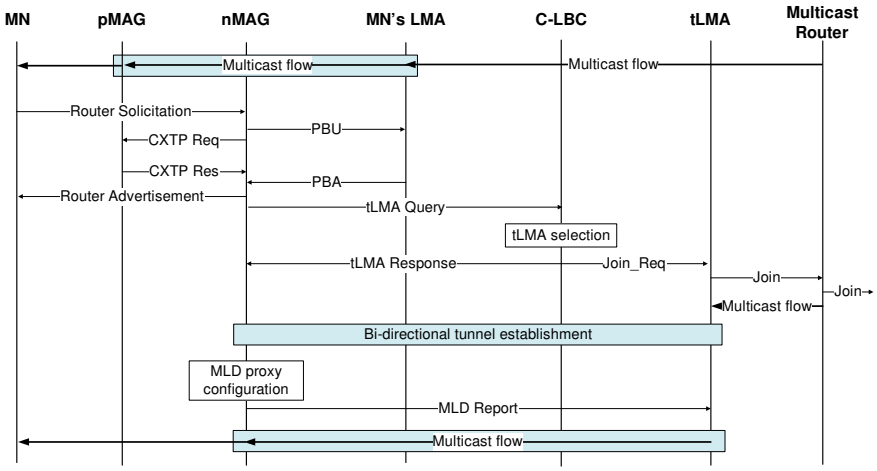


Figure 4: Handover signaling consideration with LB.

L_{lma}^{hal} depend on all the unicast sessions of the registered MNs, and the new MN arrival rate (λ_n), respectively. While L_{lma}^{lm} depends on the number of registered MNs (n) and the new MN arrival rate (λ_n). Hence, they are given by

$$L_{lma}^{mor} = \sum_{i=1}^{n}\sum_{j=1}^{u_i} L_{mn_i}^{j}, \quad (2)$$

$$L_{lma}^{hal} = \lambda_n L_{hal}, \quad (3)$$

$$L_{lma}^{lm} = (n+\lambda_n) L_{lm}, \quad (4)$$

where $L_{mn_i}^{j}$ is the load offered by the unicast flow j of the MN_i; L_{lm} and L_{hal} are the unit load generated when the LMA performs the location management and HNP allocation for an MN.

Regarding the multicast router role, L_{lma}^{mr} can be split into three main contributions corresponding to three functions: packet replication (L_{mr}^{pr}), reverse path forwarding (RPF) recalculation (L_{mr}^{rpf},) and state maintenance function (L_{mr}^{sm}) [16]. L_{mr}^{pr} is the total load from all the multicast channels which are available at the LMA, and defined as

$$L_{mr}^{pr} = \sum_{i=1}^{m} L_{mc_i}, \quad (5)$$

where L_{mc_i} is the load of channel MC_i. Note that the multicast router can replicate the data for multiple outgoing interfaces with almost the same level of load compared to that for one interface (or the unicast traffic with the same characteristics e.g., packet size and data rate) [16].

Let us now consider the different load factors which can be used as the parameters to select the appropriate LMA such as: processor capacity (CPU), number of supported sessions, number of registered MNs, and bandwidth. Accordingly, we assign each factor with a weighting variable which reflects the selected load factors. We then obtain

$$L_{lma}^{(.)} = \alpha\,(n+\lambda_n)\,L_{lm} + \beta\lambda_n L_{hal} + \gamma \sum_{i=1}^{n}\sum_{j=1}^{u_i} L_{mn_i}^{j} + \delta L_{mr}^{rpf} + \theta L_{mr}^{sm} + \rho \sum_{i=1}^{m} L_{mc_i}. \quad (6)$$

where $\alpha, \beta, \gamma, \delta, \theta,$ and ρ are weighting factors (in the interval [0,1]). For example, if the load is defined as the number of registered MNs, only two factors L_{lma}^{lm} and L_{lma}^{hal} are taken into account. In this case, the values of $\gamma, \delta, \theta, \rho$ should be set to 0. LMA load is given as

$$L_{lma}^{(.)} = \alpha\,(n+\lambda_n)\,L_{lm} + \beta\lambda_n L_{hal}. \quad (7)$$

As a result, the impact of the number of sessions as well as the session's data rates on the LMA load are ignored. Similarly, if the load is considered as the number of sessions, L_{lma}^{mor} and L_{mr}^{pr} are taken into consideration, in which the load of each session is identical. Thus, $\alpha, \beta, \delta,$ and θ should be set to 0. Eg. (6) becomes

$$L_{lma}^{(.)} = \gamma \sum_{i=1}^{n}\sum_{j=1}^{u_i} L_{mn_i}^{j} + \rho \sum_{i=1}^{m} L_{mc_i}. \quad (8)$$

Again, the impact of the session's data rate is ignored. However, it is obvious that a high data

rate session puts much more load on the LMA than the low data rate one. Therefore, they cannot be treated equally. In this chapter, we consider the sessions with different characteristics have different impact on the load.

In order to evaluate the load distribution among LMAs in different approaches, we use Jain's Fairness Index [36]. Let L denote the set of LMAs in the domain: $L = \{LMA_1, .., LMA_l\}$, where l is the number of LMAs. According to [36], the fairness index can be computed by

$$FI = \frac{(\sum_{i=1}^{l} L_{lma}^{(i)})^2}{l \cdot \sum_{i=1}^{l} (L_{lma}^{(i)})^2}, \quad (9)$$

where $L_{lma}^{(i)}$ is the load of the LMA_i (i=1,..,l). The fairness index ranges from $\frac{1}{l}$ to 1, in which the higher index indicates more fair situation. Ideally, when the load is equally distributed among LMAs, the fairness index is 1.

4.1.2. Reactive-MN and Reactive-Multicast Comparison

In the reactive-MN approach, the overloaded LMA selects an MN (say MN_{i0}) to move to a less loaded one. Therefore, the load reduction at the overloaded LMA is calculated as

$$L_r^{(mn)} = \alpha L_{lm} + \gamma \sum_{j=1}^{u_{i0}} L_{mn_{i0}}^{j} + \rho \sum_{i=1}^{m} r_i L_{mc_i}, \quad (10)$$

where

$$r_i = \begin{cases} 1 & \text{If } MN_{i0} \text{ is the last member of channel MC}_i, \\ 0 & \text{otherwise.} \end{cases}$$

In the worst case, the tLMA should join all the ongoing multicast channels of the MN_{i0}. Thus, an additional load ($L_a^{(mn)}$) is added to the tLMA load.

$$L_a^{(mn)} = \alpha L_{lm} + \gamma \sum_{j=1}^{u_{i0}} L_{mn_{i0}}^{j} + \rho \sum_{i=1}^{m} a_i L_{mc_i}, \quad (11)$$

where

$$a_i = \begin{cases} 1 & \text{If } MN_{i0} \text{ is subscribed to the channel MC}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The difference between the added and the reduced load can be considered as the waste load adding to the system. Thus, the waste load is given by

$$L_w^{(mn)} = L_a^{(mn)} - L_r^{(mn)} = \rho \sum_{i=1}^{m} w_i L_{mc_i}, \quad (12)$$

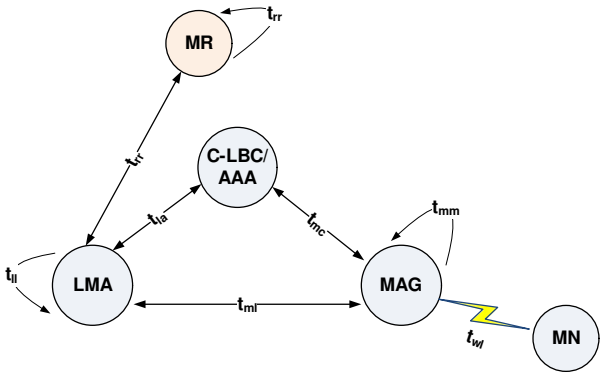


Figure 5: Reference network topology.

where $w_i = a_i - r_i$. Thus, we have

$$w_i = \begin{cases} 1 & \text{If } MN_{i0} \text{ is subscribed to } MC_i \text{ and number of listeners subscribed to this channel at the overloaded LMA is greater than 1;} \\ 0 & \text{Otherwise.} \end{cases}$$

It is noted that this value should be taken into account when selecting MN, for example, a lower value indicates more effective selection.

In the reactive-multicast approach, if the MC_{i0} is selected, the released load at the LMA is

$$L_r^{(mc)} = \rho L_{mc_{i0}}. \quad (13)$$

Following the LMA selection policy, it is high probability that the MC_{i0} is already available at the tLMA. As a result, the tLMA does not need to join this channel ($L_a^{(mc)} = 0$). In this case, the total load of the system is reduced (reduction amount $= L_{mc_{i0}}$). In other words, $L_w^{(mc)} = -L_{mc_{i0}}$. Even in the worst case, if the tLMA needs to join the channel, there is no waste load ($L_w^{(mc)} = 0$).

4.2. Multicast Service Disruption Consideration

In the reactive-MN and the reactive-multicast approach, the changing LMA of an MN (listener) may cause the service disruption of the ongoing multicast sessions. The multicast service disruption time is defined as a period when a listener cannot receive the multicast packets.

Fig. 5 shows a reference topology for performance analysis. The delay factors consisting of the total delay are defined as follows:

- t_{mm}: the delay between two MAGs.
- t_{ml}: the delay between MAG and LMA.

- t_{mc}: the delay between MAG and C-LBC.
- t_{la}: the delay between LMA and AAA/C-LBC.
- t_{ll}: the delay between two LMAs.
- t_{rr}: the delay between two MRs (between LMA and MR).
- t_{wl}: the delay between MAG and listener (MN) (wireless connection).
- t_{join}: the delay time an MR needs to join a multicast channel (including processing time and PIM Join transmission time).
- t_{qrd}: the query response delay which is the interval between the moment when the MN receives an MLD Query and replies with a report [30].
- t_{cv}: the routing convergence time which reflects the time to update the new anchor location of the selected MN's prefix.

Assuming that the delay associated with the message processing in the network entities (e.g., time for PBU/PBA processing) is included in the total value of each variable. In the reactive-MN approach, as can be seen in Fig. 1, the service disruption time (SD) can be calculated from the moment when the cLMA sends a PBU to the tLMA until the moment when the MN receives the first multicast packet from the tLMA. Let d_{join} and $d_{delivery}$ denote the time needed for the tLMA to join and get the first multicast packet for this channel (from a router which already had the multicast forwarding state for this group, namely intersection MR or IMR), respectively. Assuming that n_{mr} is the average number of hops between tLMA and IMR, we have

$$d_{join} = n_{mr}t_{join}, \tag{14}$$

$$d_{delivery} = n_{mr}t_{rr}. \tag{15}$$

Thus, the service disruption time in the reactive-MN approach is given by

$$SD_{R\_MN} = 2t_{ll}+3t_{ml}+3t_{wl}+t_{qrd}+n_{mr}t_{join}+n_{mr}t_{rr}+t_{cv}. \tag{16}$$

Via the utilization of the peering function (PF) in the reactive-MN approach, the time needed for the MLD proxy instance at the MAG to obtain the multicast subscription information can be ignored. Consequently, the service disruption can be calculated as

$$SD_{R\_MN\_PF} = 2t_{ll}+3t_{ml}+t_{wl}+n_{mr}t_{join}+n_{mr}t_{rr}+t_{cv}. \tag{17}$$

Similarly, the service disruption time in the reactive-multicast approach is computed from the moment when the cLMA sends a load warning message to the MAG until the moment when the MN receives the multicast traffic (see Fig. 3).

$$SD_{R\_M} = max\{2t_{ml}, n_{mr}t_{join}+n_{mr}t_{rr}\}+t_{ml}+t_{wl}. \tag{18}$$

Also, as seen in Fig. 4, the service disruption during handover (multicast handover latency) when applying the multicast-based LB mechanism is expressed as

$$SD_{HO} = t_{l2}+2t_{wl}+max\{2t_{ml}, 2t_{mm}\}+t_{mc}+t_{ml} + max\{t_{mc}+t_{ml}, t_{la}+n_{mr}t_{join}+n_{mr}t_{rr}\}. \tag{19}$$

where t_{l2} is the layer 2 handover latency.

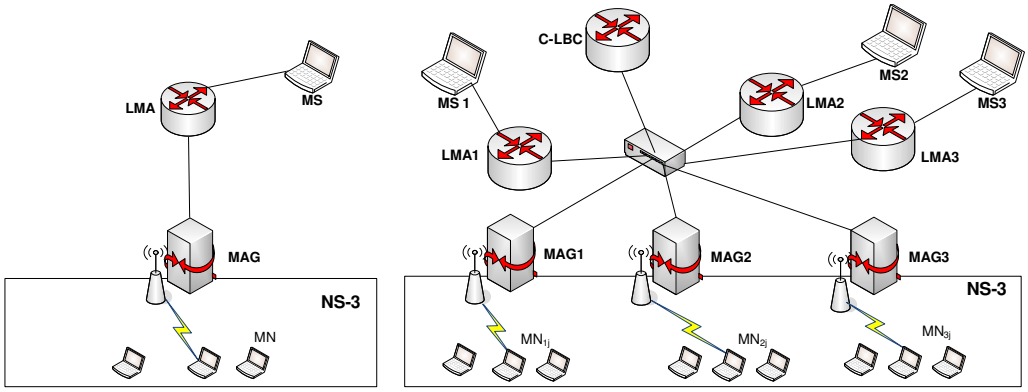
5. Experiment and Numerical Results

From the LB perspective, this section will present two separate experiments. At first, we will show in general how the different factors affect the load of an LMA. We will then evaluate the performance of the multicast-based solution in comparison with the MN-based solution and the pure-PMIP environment (without any load balancing mechanism) by using a near-to-real testbed. It is noted that, at this stage, we only focus on the case where the traffic is dominated by the multicast traffic. In addition, the load is defined as the CPU utilization rate and the performance metric is the load distribution among the LMAs. From the multicast perspective, this section will present the numerical results for the service disruption time analysis given in the previous section.

5.1. Testbed Deployment and Scenarios Description

5.1.1. Testbed Deployment

As illustrated in Fig. 6, the testbed is a combination of a virtualized environment which consists of the multiple virtual machines (e.g., using User-mode Linux) and the Network Simulator NS-3 as similar as in [37]. The PMIP entities (LMA, MAG) and the multicast sources (MSs) are the virtual machines while the access points (APs) and MNs (which play the role of a multicast listener)



(a) Experiment 1 (b) Experiment 2

Figure 6: Evaluation Testbeds.

are NS-3 nodes. More precisely, a PMIPv6 domain is deployed using an open source PMIP namely OAI PMIP [38]. Multicast support is enabled in this domain by deploying multicast router functions at LMA (using MRD6 [39]) and MLD proxy functions at MAG (using ECMH [40]). The multicast traffic is generated by a traffic generator tool (like Iperf [41], or MINT [42]). Thanks to the virtualization technique, this testbed helps to achieve the realistic results and supports a large number of MNs at a low cost. For instance, the testbed is deployed on a single physical machine even with a very limited capacity: CPU Intel Core 2 Duo T7500 (2.2 GHz) with 2 GB of RAM and 320 GB of hard disk running Ubuntu 10.04 LTS. During the experimentation, the LMA load is collected by using a performance measurement tool e.g., *mpstat*[3].

5.1.2. Impact of Different Load Factors

To show the impact of different factors on the LMA load, the first experiment used a testbed composing of one LMA, one MAG (and one AP), and one MS, as described in Fig. 6(a). Then two experiment scenarios are defined as follows:

- Scenario 1: The scenario 1 aims at demonstrating the case where the load takes into account only the number of MNs (without any user-generated traffic). The number of MNs associated with the LMA will be varied from 1 to 150 (Due to the limitation of the testbed, it can only support upto 150 MNs). The binding registration signaling for these MNs occurred within a small interval (50s) which almost represents the worst case scenario.

- Scenario 2: This scenario shows the impact of unicast/multicast flow with different data rates on the LMA load. Thus, only one MN is required. At first, the MN subscribes to a multicast channel broadcasting by the MS. The LMA load will be measured when the flow's data rate is varied from 100 Kbps to 15 Mbps. Note that a standard definition video streaming typically runs at 3.75 Mbps while the high definition at 15 Mbps [43]. The multicast flow is then replaced by the unicast one with the same data rate. The datagram size in both cases is kept constant at 67 bytes.

In order to improve the credibility of the experiment results, the LMA load was collected each one second during 360 seconds in each experiment.

5.1.3. Evaluation of the Multicast-based LB Mechanism

The second experiment aims at evaluating the performance of the multicast-based solution in comparison with the MN-based and the pure-PMIP environment. At this stage, the experiment focuses on the case where the traffic is dominated by the multicast traffic. The performance evaluation metric is the load distribution among LMAs. This metric is selected since we could not achieve high system performance without fairly and efficiently utilizing the available network resources. The other metrics such as queuing delay and packet dropping probability will be left for future works.

[3]http://linuxcommand.org/man_pages/mpstat1.html

As illustrated in Fig. 6(b), the testbed is composed of one LBC, three LMAs, three MAGs (and three APs), three MSs, and 18 MNs. The C-LBC functionality is implemented by extending the LMA functionality. At the beginning, each multicast source MS_i (i=1,2,3) broadcasts six multicast channels C_{ij} (j=1,..,6) with identical traffic characteristics (400 Kbps). In the experiment, we use the same threshold value for all LMAs, for example, 85 percent of the CPU utilization rate. At first, the MN_{ij} attaches to the MAG_i and the LMA_i, respectively. The unicast flow is also created between each MN and the corresponding MS (100 Kbps). Two scenarios (scenario 3 and scenario 4) are then defined to evaluate the proactive-multicast and the reactive-multicast approach.

In the scenario 3, six MN_{1j} (j=1,..,6) join six multicast channels C_{1j} (via LMA_1); MN_{21} joins C_{21} (via LMA_2); MN_{31} and MN_{32} join C_{31}, C_{32} (via LMA_3), respectively. Three approaches are considered: the pure-PMIP, the proactive-MN and the proactive-multicast. In the scenario 4, six MN_{ij} (j=1,..,6) join three multicast channels (say C_{i1}, C_{i2}, C_{i3}) at the LMA_i (i=1,2,3) (two MNs per channel, three channels at each LMA). Then the data rate of the existing multicast sessions as well as the number of sessions are varied to make the LMA load changes. For instance, at the LMA_1 the data rate of the channel C_{11} and C_{12} is increased with 800 Kbps and 1.2 Mbps, respectively. The channel C_{21} (at LMA_2) and the channels C_{31}, C_{32} (at LMA_3) are terminated. The results then are collected when the pure-PMIP, the reactive-multicast and the reactive-MN approach are applied.

5.2. Experimental Results

5.2.1. Load Factors Measurement

Fig. 7 reports the average and standard deviation values of LMA load as a function of the number of MNs (scenario 1). In this case, the load is calculated according to Eq. (7). We also measure the load from background processes: (average, standard deviation) = (1.001%, 0.888%). We can observe that the load slightly increases when the number of MNs increases. Fig. 8 illustrates the LMA load when the data rate of the multicast and unicast flow is varied (scenario 2). When the flow's data rate is low, the load imposed by the multicast and unicast flow is almost the same. As the flow rate increases, the load offered by the multicast flow is higher than that by the unicast flow. As

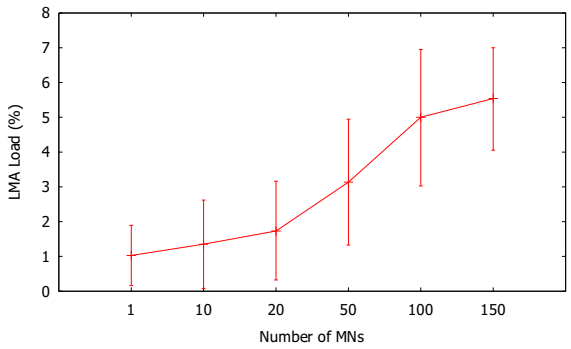


Figure 7: Load versus number of MNs (scenarion 1, experiment 1).

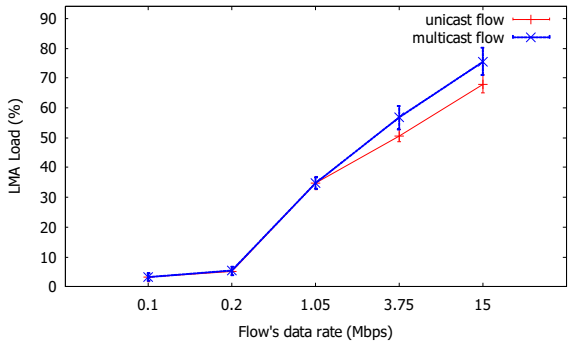
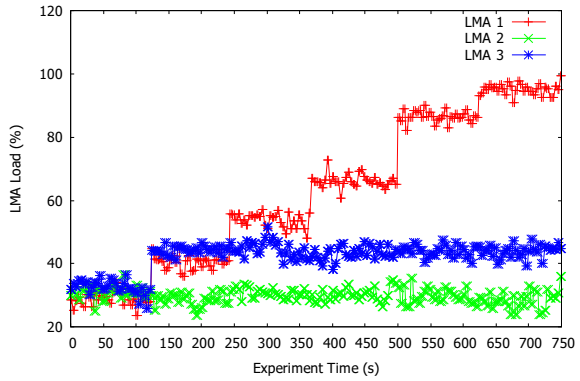


Figure 8: Load versus flow's data rate (scenarion 2, experiment 1).

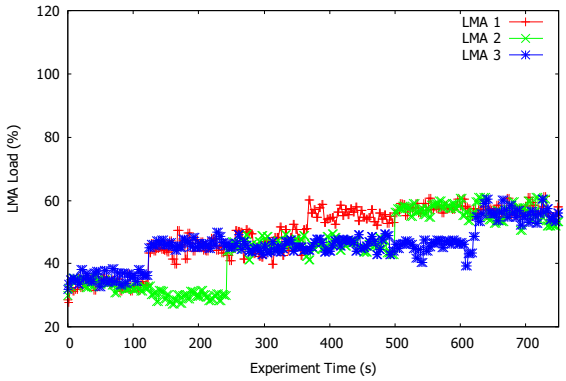
the experiment was conducted by using a very limited capacity machine, it requires about 75% load to treat a high definition video flow (15 Mbps). It also could be observed that the load offered from a typical LMA's task with 150 MNs is similar to that from a low rate multicast flow (about 200 Kbps). Thus, it is obvious that the multicast/unicast flow is a crucial factor in terms of load put on the LMA. In other words, in a multicast-dominated domain, moving an MN from the overloaded LMA could not help reduce its load significantly.

5.2.2. Evaluation of the Multicast-based Load Balancing Solution

Fig. 9 shows the FI value in the scenario 3. At the beginning, the load of all LMAs are almost the same. As a result, the FI value is very close to 1 (indicating that the load is almost shared among the LMAs). From the time the MNs subscribed to the multicast channels (at about 120s), the FI value is decreased rapidly in the pure-PMIP environment since the load is concentrated on LMA_1. For in-



(a) pure-PMIP and proactive-MN approach



(b) proactive-multicast approach

Figure 10: LMA load in the scenario 3 (experiment 2).

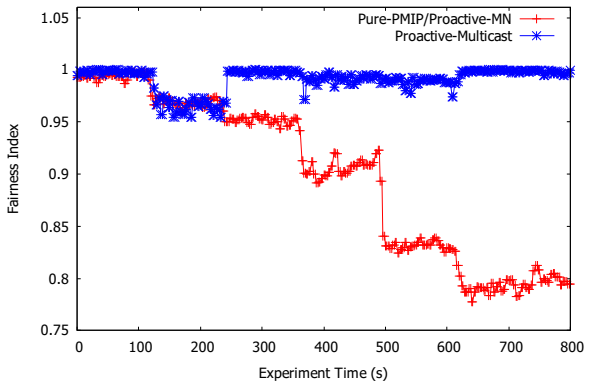
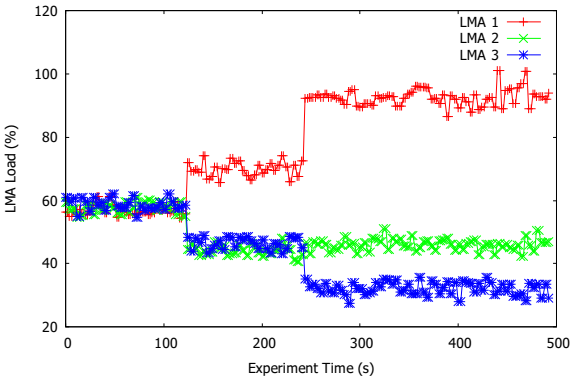


Figure 9: Fairness Index in the scenario 3 (experiment 2).

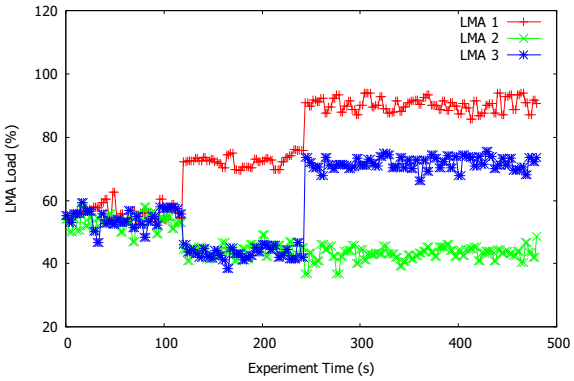
stance, LMA_1 becomes overloaded while LMA_2 and LMA_3 are at low load, as shown in Fig. 10(a). Since the LMA assignment is already done for the MNs, the FI value in the pure-PMIP can also be considered as that in the proactive-MN. It is clearly seen that the FI value in the multicast-based approach is always greater than that in the other cases (Also, the FI value is close to 1). It demonstrates that the multicast-based approach achieves a better load distribution among the LMAs. The reason is the proactive-multicast approach dynamically assigns the channel to the least loaded LMA at the time when the channel is started. In more details, the LMA load in the proactive-multicast is illustrated in Fig. 10(b). Note that the curve is not smooth since the results are collected from the near-to-real testbed (in which some background processes are generating a minor amount of load).

Fig. 11 plots the FI value in the scenario 4. At the beginning (from 0 to 120 ms), when each LMA has to serve three identical channels, the LMAs' load is nearly equal. As a result, the FI value in three approaches is almost the same and very close to 1. As the data rate of the existing multicast flow at LMA_1 is increased (C_{11}'s data rate is increased from 400 Kbps to 800 Kbps), LMA_1 load is increased accordingly. Meanwhile, the load of LMA_2 and LMA_3 is decreased (channel C_{21} at LMA_2 and C_{31} at LMA_3 are terminated). Consequently, the FI value is decreased. Since the reactive LB mechanism is only evolved when the LMA load exceeds the threshold value (85%), the FI values in three approaches are kept the same when the LMAs are running under a heavy load. When LMA_1 is overloaded (at about 240 s, as C_{22}'s data rate is increased from 400 Kbps to 1.2 Mbps), the LB mechanism is executed. As a result, the FI value in the reactive-MN and reactive-multicast is clearly greater than that in the pure-PMIP environment. That means the load is better shared between the LMAs. Moreover, the reactive-multicast approach gives a better performance than the MN-based (FI value is greater). In more details, the multicast channel with the highest data rate (C_{12} with 1.2 Mbps) is moved from LMA_1 to LMA_3 in the reactive-multicast approach, while one MN (among two) subscribed to this channel is moved to LMA_3 in the reactive-MN approach.

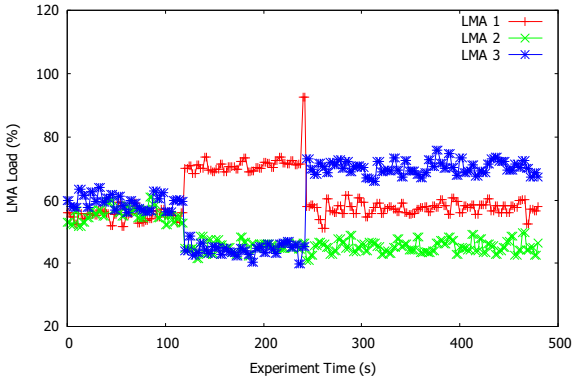
The details of load distribution of the different approaches is plotted in Fig. 12. Note that the reactive-MN helps avoid LMA_1 from being overloaded. Meanwhile, in the reactive-MN approach the overload status cannot be resolved (LMA_1 is still overloaded, while LMA_3 load is greatly in-



(a) pure-PMIP approach



(b) reactive-MN approach



(c) reactive-multicast approach

Figure 12: LMA load in the scenario 4 (experiment 2).

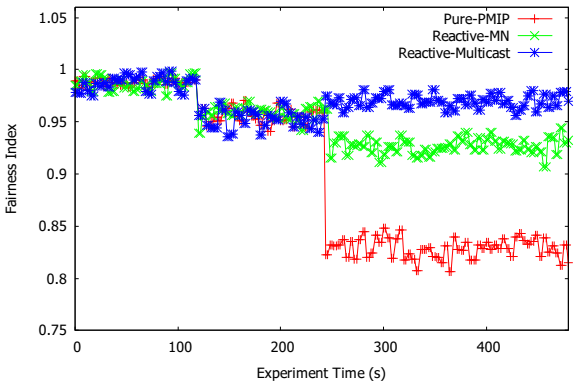


Figure 11: Fairness Index in the scenario 4 (experiment 2).

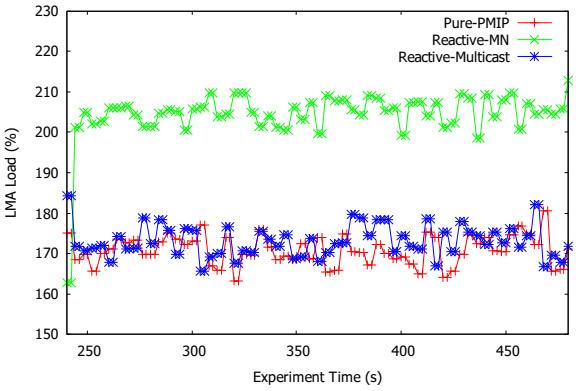


Figure 13: Total load of all LMAs.

creased). As a result, the total load of all LMAs is significantly increased compared to the pure-PMIP and the reactive-multicast approach, as shown in Fig. 13. It is due to the fact that LMA_3 has to join the channel C_{12} while LMA_1 continues forwarding this channel. In this case, more than 31% of the LMA capacity is wasted.

5.3. Multicast Service Disruption Time

In this subsection, the following parameter values are used: $t_{mm} = t_{ll} = t_{la} = t_{rr} = 10$ ms, $t_{ml} = t_{mc}$ =20 ms, t_{wl}=15 ms, t_{join} =13.5 ms, $t_{l2} = 50$ ms, and $t_{qrd} = 374.2$ ms. t_{cv} is typically in seconds (for example, the default value in case of using the Open Shortest Path First (OSPF) is 10 seconds) [44, 45]. In this subsection, it is set to 1s. The value of n_{mr} is varied over a range [0, 10] hops. It is noted that most parameters used in this evaluation are set to the typical values found in [37] and [46].

Fig. 14 shows the multicast service disruption time as a function of n_{mr}. It appears clearly that the service disruption in the reactive-MN (D_{R_MN} and $D_{R\_MN\_PF}$) is definitely higher than the maximum tolerant interruption time for normal services, as specified in [19] is 500ms. Thus, it causes a noticeable service disruption. On the other hand, the service disruption in the reactive-multicast is kept below the value of 300ms, thus, satisfying the requirements for the real-time services [19]. In other words, the reactive-multicast approach helps greatly reduce the service disruption compared to the reactive-MN solution. Moreover, in the reactive-multicast approach, if there exist the LMA which already had the forwarding state for this channel and is not overloaded, it should be chosen as the tLMA. As a result, it is high probably that the d_{join} and $d_{deliver}$ are ignored. That means, in most cases, D_{R_M} is 75 ms.

Fig. 15 shows the service disruption time during handover as a function of n_{mr}. We could observe that when $n_{mr} < 6$, the handover latency is below the value of 300 ms. Moreover, in most cases the multicast traffic is already available at the tLMA, thus, the service disruption during handover is 200 ms. Consequently, the handover impact on the quality of the multicast flow is almost imperceptible.

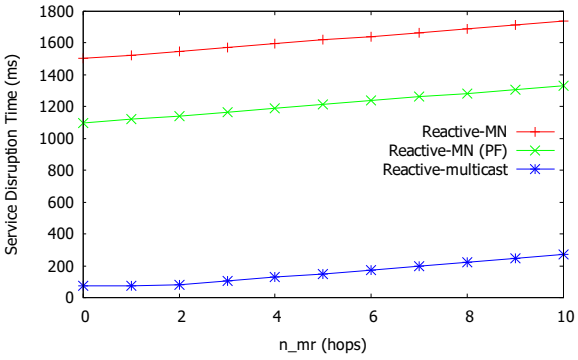


Figure 14: Service disruption time as a function of n_{mr}.

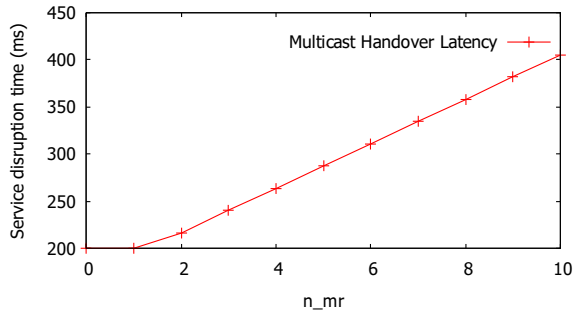


Figure 15: Multicast handover latency as a function of n_{mr}.

6. Discussions

From the performance analysis and the experiment result, we conclude that none of the two solutions are complete. The multicast-based solution in general works well in the domain where the mobile data traffic is dominated by the multicast traffic; the unicast-based solution, in contrast, works well with the unicast-dominated domain. For instance, the multicast-based solution may be the most convenient for distributing load among the dedicated multicast LMA (M-LMA) which work as the unique mobility anchor for the multicast traffic to all the registered MN in the PMIPv6 domain [18]. It comes from the fact that the M-LMA only serves the multicast traffic. As a result, the multicast-based should co-operate with the MN-based solution to enhance the reliability and scalability of the network. For example, the proactive-MN can be applied when an MN enters the PMIPv6 domain, while the proactive-multicast is evolved when a multicast session is initiated. Whenever an LMA is overloaded, if it is possible (if there exists at least one on going multicast flow), the reactive-multicast approach should be firstly performed. Then, if the overloaded state still exists, the reactive-MN should be executed. The main idea is that we try to distribute the load among LMAs by using the multicast-based solution before applying the reactive-MN solution to avoid the influence on the ongoing sessions. Therefore, the blocking probability of a new MN (session) and the dropping probability of the existing MNs (sessions) are obviously lower than the existing LB mechanisms (lower is better) .

Regarding the security issue, in our paper, it is assumed that the LMAs and MAGs which participate in the LB mechanism have an adequate prior agreement and trust relationships between each other e.g., using IPsec security association. Moreover, the tunnel between MAGs and LMAs can be pre-established as described in [10].

7. Conclusions and Perspectives

As the multicast is expected to be widely used in the future networks, degrading the role of the multicast in the available LB mechanisms can cause some issues not only from LB perspective (degradation of efficiency) but also from multicast perspective (tunnel convergence problem and service disruption). To overcome these issues, a multicast-based solution has been proposed. The benefit of the proposed solution is that it does not influence the other ongoing unicast/multicast sessions. It can also co-operate with the existing LB proposals to improve the performance of the network.

Via a near-to-real testbed, the experiment results show that the proposed solution helps better distribute the load imposed by the multicast flows among LMAs. Additionally, it helps greatly reduce the multicast service disruption time caused by the changing LMA compared to the existing proposals, even satisfying the service disruption requirement for the real-time services.

References

[1] Cisco White Paper, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017", Feb 2013.

[2] S. Kurnia, H. Lee, and S. Yang, "Understanding Consumers' Expectations of Mobile Data Services in Australia", ICMB 2007, Jul 2007.

[3] Cisco White Paper, "Cisco VNI Service Adoption Forecast, 20122017", 2013.

[4] Morgan Stanley Blue Paper, "Tablet Demand and Disruption: Mobile Users Come Of Age", Feb 2011.

[5] A. Smith, "Mobile access 2010", Pew Research Center, 2010.

[6] Ericsson, "Ericsson Mobility Report: On the Pulse of the Networked Society", Jun 2013.

[7] C. Makaya and S. Pierre, "An Architecture for Seamless Mobility Support in IP-Based Next-Generation Wireless Networks", IEEE Transactions on Vehicular Technology, Vol. 57, No. 2, pp. 1209-1225, 2008.

[8] C. Perkins, D. Johnson, and J. Arkko, "Mobility Support in IPv6", RFC 3775, Jul 2011.

[9] H. Soliman, C. Castelluccia, K. ElMalki, and L. Bellier, "Hierarchical Mobile IPv6 (HMIPv6) Mobility Management", RFC 5380, Oct 2008.

[10] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "Proxy Mobile IPv6", RFC 5213, Aug 2008.

[11] H. Yokota, K. Chowdhury, R. Koodli, B. Patil, and F. Xia, "Fast Handovers for Proxy Mobile IPv6", RFC 5949, Sept 2010.

[12] J. Korhonen, S. Gundavelli, H. Yokota, and X. Cui, "Runtime LMA Assignment Support for Proxy Mobile IPv6", RFC 6463, Feb 2012.

[13] H. Kong, S. Oh, M. Kim, and H. Choo, "Load balancing of local mobility anchors in proxy mobile IPv6 networks", Internetware '10, 2010.

[14] J. Korhonen and V. Devarapalli, "Local Mobility Anchor (LMA) Discovery for Proxy Mobile IPv6", Feb 2011.

[15] S. Jeon, R. Aguiar, and N. Kang, "Load-balancing PMIPv6 Networks with Mobility Session Redirection", IEEE Communications Letters, Vol. 17, Issue 4, Apr 2013.

[16] B. Williamson, "Developing IP Multicast Networks", Cisco Press, 1999.

[17] T. Schmidt, M. Waehlisch, and G. Fairhurst, "Multicast Mobility in Mobile IP Version 6 (MIPv6): Problem Statement and Brief Survey", RFC 5757, Feb 2010.

[18] LM. Contreras, CJ. Bernardos and S. Ignacio, "On the efficiency of a dedicated LMA for multicast traffic distribution in PMIPv6 domains", Fifth ERCIM Workshop on eMobility, Jun 2011.

[19] 3GPP, TR 25.913, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)", Release 9, Dec 2009.

[20] T.T. Nguyen and C. Bonnet, "Load Balancing Mechanism for Proxy Mobile IPv6 Networks: An IP Multicast Perspective", CNC workshop, ICNC 2014.

[21] Network Simulator NS-3, http://www.nsnam.org/.

[22] W. Fritsche, I. Gaurdini, "Deploying home agent load sharing in operational mobile IPv6 networks", MobiArch 2006.

[23] H. Deng, X. Huang, K. Zhang, Z. Niu, M. Ojima, "A hybrid load balance mechanism for distributed home agents in mobile IPv6", PIMRC, Sep 2003.

[24] J. Faizan, H. El-Rewini, M. Khalil, "Introducing reliability and load balancing in mobile ipv6 based networks", ACS/IEEE International Conference on Pervasive Services, 2006.

[25] H. Jiang, "Load sharing support for MAGs in Proxy Mobile IPv6", IETF-Draft (expired), Dec 2011.

[26] M. Kim and S. Lee, "Load balancing and its performance evaluation for layer 3 and IEEE 802.21 frameworks in PMIPv6-based wireless networks", Wireless Communications and Mobile Computing, Vol. 10, Issue 11, pp. 1431-1443, 2010.

[27] H. Kong, Y. Jang, H. Choo, "An Efficient Load Balancing of Mobile Access Gateways in Proxy Mobile IPv6 Domains", ICCSA, Mar 2010.

[28] T. Schmidt, M. Waehlisch, and S. Krishnan, "Base Deployment for Multicast Listener Support in PMIPv6 Domains", RFC 6224, Apr 2011.

[29] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, Aug 2006.

[30] B. Fenner, H. He, B. Haberman, and H. Sandick, "IGMP/MLD-Based Multicast Forwarding (IGMP/MLD Proxying)", RFC 4605, Aug 2006.

[31] T. Schmidt, S. Gao, H. Zhang, and M. Waehlisch, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains", IETF-Draft (work in progress), Oct 2013.
[32] H. Zhang, et al., "Multiple Upstream Interfaces IGMP/MLD Proxy", IETF-Draft (work in progress), Jul 2013.
[33] V. Devarapalli, R. Koodli, H. Lim, N. Kant, S. Krishnan, and J. Laganier, "Heartbeat Mechanism for Proxy Mobile IPv6", RFC 5847, Jun 2010.
[34] F. Xia, B. Sarikaya, J. Korhonen, S. Gundavelli, and D. Damic, "RADIUS Support for Proxy Mobile IPv6, RFC 6572, Jun 2012.
[35] S. Krishnan, et al., "Update Notifications for PMIPv6", IETF-Draft (work in progress), Aug 2013.
[36] R. Jain,"Throughput fairness index: An explanation", 1999.
[37] T.T. Nguyen and C. Bonnet, "Performance Optimization of Multicast Content Delivery in a Mobile Environment based on PMIPv6", WCNC 2013, Apr 2013.
[38] OAI PMIPv6, http://www.openairinterface.org /openairinterface-proxy-mobile-ipv6-oai-pmipv6.
[39] MRD6, http://fivebits.net/proj/mrd6.
[40] ECMH Easy Cast du Multi Hub, http://unfix.org/projects/ecmh/.
[41] Iperf, http://sourceforge.net/projects/iperf/.
[42] MINT, http://mc-mint.sourceforge.net/.
[43] D. Smith, "IP TV Bandwidth Demand: Multicast and Channel Surfing", INFOCOM 2007.
[44] Y. Tsegaye, T. Geberehana, "OSPF Convergence Times", Master Thesis, Chalmers University of Technology, Goteborg, Sweden, 2012.
[45] Configuring OSPF Timers, http://www.juniper.net/techpubs/en_US/junos11.4/topics/topic-map/ospf-timers.html.
[46] S.J. Yang and S.H. Park, "A Dynamic Service Ranged-Based Multicast Routing Scheme Using RSVP in Mobile Networks", GLOBECOM, 2001.

Tien-Thinh Nguyen received the B.S. degree in Computer Engineering from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2006. He received the M.S. degree in Networks and Communication Systems from Hanoi University of Science and Technology, Hanoi, Vietnam and University of Claude Bernard Lyon 1, Lyon, France in 2010. He is currently a Ph.D. candidate in the Department of Mobile Communications at EURECOM, Sophia Antipolis, France. His research interests are in the areas of wireless networks and mobile networks with emphasis on IP multicast and IP mobility.

Christian Bonnet received his M.S. degree in 1978 from Ecole Nationale des Mines de Nancy, France. He is currently a Professor in the Department of Mobile Communications at EURECOM, Sophia Antipolis, France. His teaching activities are real-time and distributed systems, mobile communication systems, wireless LANs and protocols for mobility management. His main areas of research are wireless protocols, wireless access to IP networks and data communications in mobile networks including mobile ad hoc networks.